Review of: "An Improved Hybrid Transfer Learning-Based Deep Learning Model for Alzheimer's Disease Detection Using CT and MRI Scans"

Steven Frank

Potential competing interests: No potential competing interests to declare.

The authors describe an interesting study directed toward an important clinical challenge. The article is well-written and timely, but I suggest some clarifications and further explanation.

The title of the article is: An Improved Hybrid Transfer Learning-Based Deep Learning Model for Alzheimer's Disease Detection Using CT and MRI Scans. It seems, however, that the authors simply tested various CNN architectures, none particularly new, on the dataset separately. So there do not seem to be any improvements, nor is anything "hybrid." It appears the authors believe that using pre-trained (ImageNet) weights constitutes "transfer learning," but most practitioners would disagree; transfer learning generally uses models first trained on a related domain, and ImageNet is not related to AD. Finally, the title refers to CT and MRI scans, but it appears that their dataset only involves MRI.

This points to a broader deficiency in the presentation. The authors do not describe the dataset in adequate detail, nor do they specify exactly what they did or measured. As an example, the authors say, "Here in this study, we use downsampling techniques to balance the data." How does downsampling balance data? What was the degree of imbalance?

Tables 2 and 3 show five classes. Did the authors train these various CNN models to classify test data in these classes? If so, classification accuracy should be reported per class. Moreover, it is not clear what the results mean. Table 2, for example, is titled, "Classification report generated by the VGG16 model" – what does this mean? How are metrics generated? Is this for the architecture generally, or on this dataset? In reporting precision, recall, and F1 scores, what does this mean in terms of classification accuracy? How do the authors define false positives and false negatives, which underlie these metrics, for this classification problem? Tables 2 and 3 report results for VGG16 and ResNet50; what about the other architectures the authors investigated?

More generally, and perhaps most significantly, it is unclear what the authors' contribution is. Did they simply apply conventional CNN architectures to a publicly available dataset in a conventional way? Or did they add something original?